Statistical disclosure methods

*With focus on Data swapping*

Heval Beydogan and Wanmeng He


Supervisor: Dan Hedlin
Abstract

Statistical disclosure control is a method that is used to protect confidentiality of respondents. Moreover, when a data set is protected some of the information is lost and the choice of method has to be balanced against the information loss and the confidentiality of the respondents. Also, the choice of method not only based on information loss versus utility but also on current legislation. There are various methods that can be implemented when protecting confidentiality. In this paper we study a range of statistical disclosure control methods, but mainly the method called data swapping. Our aim is to analyze the advantages with data swapping compared to other methods and whether it is a commonly used method by statistical organizations. The advantage of data swapping is that no information is removed from the data set; however, the structure and some of the utility will be lost.
# Table of Contents

1 Introduction ................................................................................................................. 3  
1.2 Purpose ....................................................................................................................... 3  
1.3 Methodology ............................................................................................................... 4  
1.4 Limitations .................................................................................................................. 4  
1.5 Outline ....................................................................................................................... 4  

2. The need of disclosure control ..................................................................................... 4  
2.1 Risk ............................................................................................................................ 5  
2.2 Information loss .......................................................................................................... 6  

3. Overview ....................................................................................................................... 7  
3.2 Tabular data ............................................................................................................... 7  
3.3 Methods for tabular data ......................................................................................... 8  
3.3.1 To evaluate risk in tabular data ........................................................................... 9  
3.4 Micro data ................................................................................................................ 9  
3.5.1 Masking micro data ............................................................................................. 10  
3.5.2 Synthetic method ................................................................................................. 12  
3.5.3 Restricted access and restricted data ................................................................. 12  

4. Data Swapping ............................................................................................................. 12  
4.1 Definition .................................................................................................................. 12  
4.2 Development ............................................................................................................. 14  
4.5 Data swapping methods ............................................................................................ 15  

5. How do statistical organizations work? ...................................................................... 23  

6. Discussion and summary ............................................................................................. 25  

7. References ................................................................................................................... 27
1 Introduction

Statistical organizations publish statistical information. Their objective is that the outputs of the statistical information are rich in information while at the same time following legislation and public demand on secrecy and confidentiality. The amount of statistical information is based on the type of data publishing. Micro data contains more information than tabular data, which means that the risk of disclosure is higher with micro data. Disclosure happens when information about a respondent is revealed through published statistical information. Statistical disclosure control (SDC) is a term for methods to protect confidentiality of respondents. Given the form of output, which can be either post tabular, which is tabulated aggregated data, or pre tabular, which is micro data; there are different SDC-methods to approach the need to protect respondent confidentiality. Data swapping is one of the methods to protect micro data. The basic idea behind data swapping is that values in a data set get swapped within the data set in order to protect confidentiality.

Dalenius and Reiss (1978) first introduced the concept of data swapping. Their method is designed for categorical data. The original methodology have since been developed and altered by, among many, Moore (1996), Carlson and Salabasis (2002) and Shlomo et al (2010).

Although there are many SDC-methods each statistical agencies must deal with statistical data depending on the data, the form of the data and the variables in the data. That is, there is no clear-cut answer to which method to use. Decisions are based on type of data, how it is going to be published and an evaluation of what variables are at risk of disclosure. Our objective with this thesis is to give an overview of available SDC-methods and particularly data swapping since we believe data swapping have advantages over other techniques.

1.2 Purpose

The main purpose of this thesis is to study the statistical disclosure control method called data swapping, a technique that swaps variables in a micro data file so that personal information is not revealed. Moreover, the thesis addresses other disclosure control methods for micro data and aggregated data. The paper is mainly a literature study over the current statistical disclosure control methods. However, the aim of the paper is data swapping. Our purpose is to analyze whether data swapping adequately protects individuals and firms from disclosure and if the method is implemented by statistical organizations and how it compares to other SDC-methods.
1.3 Methodology
The thesis is based on literature studies and interviews. The literature which has been used consists of books and several scientific articles from both old publications and the most up-to-date articles. In order to get the information about how statistical organizations work for protecting data, we have interviewed staff in Statistics Sweden.

1.4 Limitations
First we give an overview of statistical disclosure control methods. However, we focus on one of these methods; data swapping.

1.5 Outline
This thesis begins with a short overview of why disclosure control is needed. The concept of information loss will be touched upon. In chapter 3, different statistical disclosure control methods are introduced. Data swapping as one of the statistical disclosure control methods is presented in chapter 4, as a topic of this thesis; several variants of data swapping are briefly explained. In chapter 5 we present how statistical organizations work to protect data, the UK census, The Australian Bureau of Statistics and Statistics Sweden are mentioned. The thesis is concluded with a discussion and a summary.

2. The need of disclosure control
The definition of disclosure is when an individual or organization is identified, or information about an individual or firm is revealed, through published statistical information. The objective is to minimize risk of disclosure that can hurt individuals or companies, and at the same time produce official statistics that have an acceptable quality. Generally disclosure control doesn’t eliminate the risk of identification; however, the aim is to find a balance between the risk of disclosure and the loss of information. EU regulation states that statistical information that is collected should protect personal information and “respect for private and family life” (Regulation (Ec) No 223/2009), in order to maintain the confidence of the respondents.

When publishing statistical information there is a tradeoff between the amount of detailed information that is published and the protection of respondent confidentiality. That is, the tradeoff is between provision of information and protection of those whom provide the information.
The aim with disclosure control is to minimize the risk of identification and to keep a high utility of the statistical information.

Furthermore, to maintain quality of statistical information it is crucial that respondents perceive that confidentiality is preserved. Confidentiality is needed for reasons such as legislation, (Swedish, European Union and United Nations), and not to violate respondent trust. According to EU regulation, data should be stored in such way that unauthorized personnel don’t have access to confidential data, physical protection, and by statistical disclosure control to ensure confidentiality.

In Sweden the ambition is to facilitate research through accessibility of register data. However, the accessibility has to be evenhanded with the protection of individual integrity. With modern technology at hand, where affordable and effective technology is more available than ever before, statistical organizations and agencies must adapt to protect against disclosure of sensitive and confidential information.

Today Sweden lacks an infrastructure that facilitates an effective use of micro data at statistical organizations. Moreover, the need for protection of people’s integrity has to be balanced to the facilitation of register and micro data that has analytical value.

2.1 Risk

Publication of statistical data poses disclosure risks. To dress potential issues that disclosure risk imposes risk analysis are made. In risk analysis of statistical data there ought to be rational assumptions of what intruders might know beforehand, in order to assess what sort of disclosure method one should use. That is, evaluating a disclosure scenario. We call anyone who reveals confidential information for “intruder”, in which people who disclose a respondent without the intention to do so, is also called “intruder”.

A disclosure scenario is when an intruder can; with or without prior information and statistical data identify individuals or statistical units. Identification can be direct, where identification is made from formal identifiers, such as name and address. Moreover, identification can be indirect, where combinations of variables, such as education, age, and religion, make identification.

What information is available to the intruder, and how can the intruder use this information to identify an individual or firm? These are some of the questions that are posed when assessing the risk scenarios. We believe that the widespread use of Internet and the availability of information
that comes with it, assessing risk scenarios are much harder than before. Perhaps in analyzing
disclosure scenarios more emphasis should be on what information people themselves publish
and information that can be gathered from web sites. Meaning, how much information can an
intruder gather from investigating on the web, and how does this affect the possibilities of
identifying statistical units.

In addition to prior information, the survey types also affect the risk scenarios. Data that is
based on census contains more information, hence more key variables that can facilitate
identification. Data that is based on surveys are somewhat self-protecting. To estimate disclosure
risk on survey data, probabilistic models are used.

Different kind of publications has different kinds of disclosure risks and different kinds of
statistical disclosure methods, which we will discuss in chapter 3.

2.2 Information loss

One of the challenges with SDC is to protect individuals and firms while keeping the information
loss at minimal. Information loss is when the masked data and the original data differ in results.
That is, if the difference is high then the information loss is high and vice versa.

Often when assessing disclosure risk the knowledge about an intruder is limited, hence the
evaluation of what to mask and the effect of information loss are hard. However, the information
loss is evaluated based on the uses of the data. Since the uses widely vary so does the impact of
information loss. Domingo-Ferrer and Torres (2001) suggests that a generic evaluation of
information loss should correspond, although roughly, to a range of data usage.

Nevertheless, it is important to estimate the harm of information loss, so that measures
whether the masked data set still has an analytical value can be made. The aim is to preserve the
structure of the data set so that it still is has analytical value. According to Winkler(1998) if the
following three conditions are approximately preserved, the data set has analytical value.

- Means and covariances on small set of sub domains.
- Marginal values for few tabulations of the data
- At least one distributional characteristic.
Nevertheless, some only apply on continuous variables. Also, if six variables on important sub domains are provided then the data is analytically interesting. The conditions are somewhat unclear.

3. Overview

If, in the risk analysis, it is found that there are risks for disclosure, the next step is to find methods to prevent identification of units. There are some alternatives to provide statistical confidentiality through statistical disclosure control, one can either protect tabular data (information that is aggregated and presented in tabular form), or protect micro data.

Moreover, there are various intruder scenarios and one has to take to account what an intruder might know beforehand and what the intruder might decipher from statistical information with that prior knowledge. A situation where information about statistical units does not need prior knowledge to be identified is called spontaneous recognition. Different kind of scenarios imposes different measures to inhibit disclosure.

A measure that minimizes disclosure also decreases the amount of information in the data. Therefore the quality of the modified data must be measured. The last step is to evaluate whether the data can be used or if there are other methods to apply.

It is vital to document what kind of statistical disclosure method has been used. However all the details need not be published, since a too detailed report can in itself help an intruder to detect sensitive or classified information about an object. Nevertheless, important information such as error terms, sample size etc. ought to be reported in the publication.

Depending on how data is published there are methods to prevent disclosure. We will give a short overview SDC-methods for tabular data and micro data.

3.2 Tabular data

Tabular data, which is aggregated data, is often presented as magnitude table or frequency table.

Magnitude table: In magnitude tables the cell values represent the sum of quantitative variables. The variables are grouped and presented in tables so that the cell values represent the sum of the
grouping of values. Table with a lot of details have many groups, and a table with less information has fewer groups. The typical respondents are businesses.

In the case of magnitude tables, the risk of disclosure occurs when the cells contain small values. Cells with small values, and hence a higher risk of disclosure, are called unsafe, confidential or sensitive cells. One of the statistical disclosure controlling methods that are used is cell suppression.

*Frequency table:* In frequency tables the cell values, which are in percentages, represent the number of persons that represent that particular cell.

When cell values are small, so called unsafe cells, the risk of disclosure is higher, as in the case of magnitude tables.

### 3.3 Methods for tabular data

*Pre-tabular methods:* methods that are used on micro data. Once SDC-methods have been used on micro data the SDC-methods for tabular data is not needed. However, access to original data is needed and the distributions in the data can be distorted. Also, this method is not as transparent as the other methods.

*Table redesign:* includes thresholds, reducing detail levels and minimum average. Original counts are not affected. This is a simple method.

*Post-tabular:* modifying the cell values in a table. Tedious since each cell must be evaluated separately. Methods are cell suppression, rounding and cell perturbation.

*Threshold method:* in frequency tables, if a cell in a table is less than a certain threshold value, then the cell is risky. Common threshold values are 3, 5 or 10. These values are not a rule but they are probably the values most organizations use. The rule is used to protect cells with small values, and to give a certain protection against disclosure when used together with other tables. However the margins in the tables can reveal information. This method is also called the minimum frequency rule.

*Cell suppression:* is used to hide or mask a cell value, and to replace the numerical value with an alternative symbol. Primary suppression is masking of an insecure cell. Secondary suppression is masking of safe cells with the purpose of hindering derivation of the primary suppressed cells. There is an optimization problem with secondary suppression since there are many alternatives. The best solution is the one with least information loss. One can have a security interval for the
primary cell, it large enough so that estimation is impossible. Secondary suppression is made with the interval at mind. If the total is published in a magnitude table a statistical object can be disclosed even if a sensitive cell has been suppressed; simply by adding all other cell values and subtracting that sum from the total. To strive this, secondary suppression is used. If the total is not published secondary suppression is not needed. Large group margins are kept.

**Rounding:** used on frequency and magnitude tables. One can either use uncontrolled rounding or controlled rounding. Uncontrolled rounding is when the sum of the cells is rounded so that the cell sums do not add up to the marginals. Controlled rounding is when the cell values are rounded so that the sums of the cell values are the same as the sum of the marginals. Rounding creates uncertainty about the real values.

### 3.3.1 To evaluate risk in tabular data

**P% rule:** is used on magnitude tables without negative values as risk assessment of data. The aim is to state a permitted value on how close calculations can get the real value. It is shown that the object with the second largest value in the cell gives the best possibility to estimate another object's value (the largest object). For instance, compute the total and subtract the two largest contributors. If the two largest contributes are less than p% of the total, disclosure control is not necessary. That is, this method is used when the disclosure scenario induce identification of the largest object in the data set and to evaluate whether data needs protection.

**N,k-rule:** also called the dominants rule. Used on magnitude tables without negative values. A cell is considered risky if the n largest objects variable value is more than k% of the cell total. A typical case is when total sum of the n largest objects is larger than k% of the total of the cell. It is also important to keep the parameters confidential. This method can be used on negative values with some modifications to the calculations. This method has a tendency to overprotect the data and is considered by some statistical organizations to be unusable.

### 3.4 Micro data

Micro data that doesn’t concern people or businesses does not need protection and can be published as it is. In other cases, estimation of who will get access to data and what information the data contains should be made.

Geographical facts about individuals and firms can be sensitive and one should be careful about how detailed published information is. Regardless, one has to differentiate between research statistics and official statistics. Statistics that researchers use is dependent on the current legislation of the country the researcher resides in. In Sweden research statistics can contain more
detailed information than ordinary published data, provided the privacy is maintained. However, official statistics, that is published, must fulfill privacy legislation.

Depending on the character or structure (hierarchical or not) of the data, evaluation and risk assessments of which statistical disclosure control method works best should be done. Moreover, contingent on the information in the data, objects that are included in the data and which of these objects that is a risk factor should be evaluated. The sample method, whether its form a register or stratified sample, questionnaire, should to be evaluated and assessed as well. Micro data can be masked, i.e. generate modified data or create a synthetic data from the original data.

3.5 Methods for micro data

3.5.1 Masking micro data

There are two categories of masking, both with effects on data. The first method is perturbative methods, which either modifies the identification variables or modifies the confidential variables. In other words, data that is falsified by introducing erroneous elements to the data. The other method is non-perturbative methods.

In the first case, unique combinations in original data disappear and are presented as other new unique combinations. In this way, users can never be sure if the uniqueness of one unit truly is a unique quality of that unit. This implies that when having identified a unit, one can never be sure if the values are true.

The second method does not change the values on the variables, but produces data with suppressed information. Nevertheless both methods have effect on information loss.

Masking is a method that facilitates publications of whole micro data sets, while containing perturbed elements. This method is used on continuous data. Below examples of masking will be presented as suggested by Duncan et al (2011).

*Masking through suppression*: To protect micro data, suppression can be performed as

(1) record suppression or
(2) attribute suppression.

Deleting data records is the way to perform record suppression. For instance, delete all records that have a weight over 90kg. Attribute suppression means that attribute values are deleted, such as deleting the attribute “wage”, which will reveal private information for data intruder.

Three cases of masking through suppression have been mentioned by Duncan et al (2011) where the disclosure risks are extremely high. These three cases can be summarized as

(1) when longitudinal data appeared in records
(2) the records provide geographical specificity and
(3) the records have hierarchical structure.
These problems can be solved by longitudinal suppression, geographic area suppression and hierarchical suppression.

*Masking through sampling* Masking through sampling can be easily explained as the statistical organizations take samples from the record and then mask selected data. Now consider if the source data itself is a sample, then the data can be defined as self-masked and other disclosure limitation methods should be used. Masking through sampling works so that a data snooper feels uncertain whether any linkage between a population unit and a data unit is correct.

*Masking through aggregation:* Masking through aggregation can be divided into global coding and top coding. To use “aggregation” as a masking method means in principle combining attribute values. The process of global coding is to combine several categories of an attribute into a single one. For example, a statistical organizations would like to combine attribute “place of residence”, then “Stockholm”, “Täby”, “Solna” “Kista” can be combined as “Stockholm region”. Moreover, global coding is not only used on data with high disclosure risk, but on the whole data file. Top coding is a specific example of global coding, where all values above a certain threshold is notated with either a summary statistic or as conditional mean or median

*Local suppression:* Local suppression is a more complicated pattern compared with the different suppression methods mentioned above. Local suppression suppresses certain values of individual attributes. (Willenborg and de Waal, 2001). In other words, consider a record containing several attributes; a set of combinations of attributes in that record poses a high disclosure risk. As an example, the following attributes combined have a high disclosure risk; “Age=37”, “Sex=man”, “residence=Kalix” and “occupation= teacher”. After applying local suppression, the combination of the above mentioned values appeared as “Age=missing”, “Sex=man”, “residence=Kalix” and “occupation= teacher” In this way, the data snooper can no longer identify the record, and the disclosure risk has been lowered.

*Noise addition:* Consider this equation $y=x+\varepsilon$, where $x$ presents attribute values and $y$ presents masked value. The mean of $\varepsilon$ is assumed to be zero, $\varepsilon$ is a random variable when added with $x$ works as a noise in order to prevent data snooper to link data to an identity linker database.

*Data swapping* is a technique to mask data files, so that enough information is contained while disclosure risk is low. This method will be more thoroughly examined in chapter 4.
**Micro aggregation**: The process of this method can be summarized as: identifying the attributes which may have high disclosure risk, after that, divide these attributes into clusters according to the attributes and conduct the aggregate values (mean, median, etc). Use these aggregated values to replace the original data, which have high disclosure risk.

### 3.5.2 Synthetic method

Unlike the methods mentioned above, this method does not need to mask source data. Masking data works directly on the source data, but synthetic data requires building a probability model statistically estimated from the source data which is called “synthesizer” and the data derived from this model works as the release data. The concept of this method is easy, but it is difficult to apply in the reality.

### 3.5.3 Restricted access and restricted data

To restrict access can be easily implied, firstly the statistical organizations can introduce licensing or bonding so the only limited individuals or organization can get access to the database. Secondly, before the researchers release their products, the statistical organizations have the right to review the product so that they can ensure that no confidential information is revealed.

The concept of restricting data means that the released data has been transformed from the source data, and these data do not have disclosure risk. The advantage of this strategy is that the released data are not only available to a limited set of people.

### 4. Data Swapping

#### 4.1 Definition

Data swapping is a disclosure control method for protecting micro data that is going to be published either as micro data or tabular data. This technique is sometimes referred to as multidimensional transformation and data switching. The objective of data swapping is to lower the disclosure risk and protect the confidentiality in released micro data. In this way, the data users or data snooper/intruder will be uncertain about whether released data correspond to real data elements.
Fienberg and McIntyre (2005) described the basic idea of data swapping as a transformation of a database by exchanging values of sensitive variables among individual records. This idea can be interpreted as using the original data matrix $M_1$ as the “input” for producing another perturbed data matrix denoted as $M_2$. In this case, $M_2$ is the data set used to calculate statistics. This process can be conducted by exchanging attribute scores between respondents. Clearly not all attributes in a record will be swapped; only certain attributes will be swapped. These attributes are defined as swapping attribute and the concept of swapping attribute is defined by Duncan et al (2011) as an attribute over which swapping is to occur. After data swapping, the value obtained by one respondent may belong to someone else; each respondent’s anonymity is protected.

Here is a simple example of how data swapping is processed. A micro data file contains 3 variables for 8 respondents. In order to protect the privacy of respondents, data swapping processed need to apply. In this example, the values of age are randomly swapped, those on record 1 and 5, and those on record 3 and 6 are pairwise swapped, the values of other variables remain the same.

Here is a simple example of how data swapping is processed. A micro data file contains 3 variables for 8 respondents. In order to protect the privacy of respondents, data swapping processed need to apply. In this example, the values of age are randomly swapped, those on record 1 and 5, and those on record 3 and 6 are pairwise swapped, the values of other variables remain the same.

<table>
<thead>
<tr>
<th>Record</th>
<th>Age</th>
<th>Geography</th>
<th>Occupation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>20</td>
<td>4</td>
<td>8</td>
</tr>
<tr>
<td>2</td>
<td>35</td>
<td>4</td>
<td>7</td>
</tr>
<tr>
<td>3</td>
<td>21</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>4</td>
<td>34</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>66</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>6</td>
<td>37</td>
<td>1</td>
<td>9</td>
</tr>
<tr>
<td>7</td>
<td>24</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>8</td>
<td>50</td>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>

Table 4.1 Example of data swapping
4.2 Development

Dalenius and Reiss first introduced, data swapping, as a disclosure control method in 1978. Their data swapping technique was mainly suggested for usage in cases where the attributes in a data set assume categorical variables. At that time data swapping techniques had not been tested on real-life data, all illustrations from their report were based on small-scale numerical experiments.

Later on, Greenberg (1987) introduced a new data swapping procedure named the rank-based proximity swap procedure. This procedure can be used on continuous variable. Unlike the original procedure, the rank-based proximity swap procedure restricts the range for which each value can be swapped. 1996, Moore published a paper with an enhanced rank-based proximity swap algorithm.

4.3 Implementation

There are some programs that can be used to mask micro data. Among all, code software packages such as SAS.

The webb based tool NISS, developed by National Institute of Statistical Science (NISS), can perform data swapping on categorical variables. The Webb tool not only swaps data but also provides a risk versus utility computation that quantifies the tradeoff between the risk of disclosure and the information loss. The usage is straightforward. A micro file is uploaded; variables (attributes) that one wants to swap and the swap rate are specified. Swap rate is the fraction of the variables that is marked to be swapped and some of the specified variables get swapped. Users are provided with an output file containing details of the process.

Another software tool that can be used is µ-argus. It is a free program provided by statistics Netherlands that is used on micro data. (The other version, τ-argus is used on aggregated data.) µ-argus needs both a micro data file and a meta-data file that describes the variables in the micro data. The program can be implemented in statistical software packages such as SAS and SPSS. µ-argus can be used for numerical rank swapping as described by Moore (1998); the swapping algorithm is independently used on each variable. Otherwise Global recording and local suppression are the two main techniques in µ-argus.
4.4 Advantages and disadvantages

Advantages

- Data swapping removes the relationship between the record and the respondent. This can be considered as a psychological advantage, since it is impossible for data user or data snooper to say if the sensitive properties assigned to a person are his own or not. (Carlson and Salabasis, 2002)

- Data swapping could be very effective with a low cost of implementation. The procedure is easy to operate, some data programs can be applied nowadays which makes it even easier than before.

- Data swapping can be used on sensitive variables without disturbing non-sensitive variables. It can provide the protection where it is most needed. This implies that data swapping procedure can ensure the accuracy of statistics as much as possible.

- There is no limitation of variables for applying data swapping, both categorical variable and continuous can be swapped.

- The information loss is relatively small compared to some post-tabular SDC-methods.

- The procedure does not disturb non-sensitive fields.

Disadvantages

- Obviously, data swapping will affect the structure of the data, and arbitrary swapping will provide inaccurate statistics. According to Moore (1996), a large number of records with unusual combinations could be produced by arbitrary categorical swaps and arbitrary continuous swaps.

4.5 Data swapping methods

General data swapping: Consider a variable/field that is regarded as a sensitive variable, income for instance. Within a micro data file, with “income” as one of the variable fields, income values are swapped within the file and field, to protect the confidentiality of respondents. Variables in other fields in the same data file can be swapped as well. For instance in a file with age, income, sex etc as variable fields, and both fields “age” and “income” can be swapped. That is the variables within each field can be swapped. The choice of what variables to swap is dependent on the perceived sensitivity of the variables. The univariate distribution and the anonymity are
protected using this method.

*t-order frequency count:* The t-order frequency count is used on categorical data sets. The term t-order refers to the number of variables, so if a data set contains two variables (x and y) it is a 2-order statistic. Variables are swapped so that a lower order frequency or marginal is preserved. The preserved frequency is pre-specified. However, this method was not been used in practice but was a theoretical incipience for the methodology data swap. The method in practice, however, changes distributions. If there is a swap between a clerk and an executive the income distribution is altered and the analysis can not only provide wrong coefficients but also provide with wrong conclusions. That is, there might occur an analytical predicament if the distribution for two variables is altered in a way that affects the analysis. This methodology is not sufficient to protect from disclosure with the current accessibility and use of data and software.

*Rank based swap:* A more modern version of the t-order frequency count is a rank based proximity algorithm introduced by Greenberg (1987). It is a method used on continuous variables. The range for the swap is restricted; hence the distortion is less than in the t-order frequency count, and the estimates are better. The suggested algorithm that follows clarifies the version Moore (1996) describes while maintaining his notation to facilitate comparison.

1. Start with a data file of size N and order responses by a single variable, a. That is, index responses to a, by i=1, 2,...,N; where a_i < a_j, if i < j. i j
2. Determine a value P(a), with 0 < P(a) < 100. The intent of the procedure is to swap the value of a_i with that of a_j, so that the percentage difference of the indices, i and j, is less than P(a) of N. That is |i - j| < P(a) * N/100.
3. Initialize all ranks with a set to a top- or bottom-code as "swapped". Also initialize the ranks of all imputed and blank values to "swapped". Initialize all other ranks as "un-swapped". That is, variable values that are coded as either top- or bottom code is not eligible for swapping. See figure 4.1.
4. Let j be the lowest un-swapped rank. Randomly select a record with an un-swapped rank from the interval [j+1, M] where M= min {N, j + (P(a)*N/100)}. Suppose the randomly selected record has rank k.
5. Swap the values a_j and a_k. Set the labels on these to ranks to "swapped".
6. Return to Step 4 and continue until all ranks are labeled "swapped".
7. Suppose one swaps on several additional fields, b, c, ... . Return to Step 1 and repeat the procedure one field at a time. First use field b, then field c, ... . P(b) need not equal P(a).
8. When the swap is complete, calculate and compare multivariate statistics. If they are not
within a suitable range, repeat the procedure using smaller values for \( P(a) \), and/or \( P(b) \),... 

To exemplify the algorithm above we chose values for \( N, a \) and \( P(a) \).

\( N = 100, \)
\( a = \text{income}, 0 < a < 100 \)
\( P(a) = 5, \) this value represents the relative distance of the swapping window.

The first step is to order the values by index. That is;

\[
\begin{align*}
a_1 &= 1 \\
 a_2 &= 2 \\
 a_3 &= 3 \\
 &\vdots \\
 a_{98} &= 98 \\
 a_{99} &= 99 \\
 a_{100} &= 100
\end{align*}
\]

The second step is to determine the value of \( P(a) \), that is the window which within the swapping will occur. The absolute difference between two swapped variables should be less than \( P(a) \times N/100 \), that is;

\[
|i - j| < 5 \times \frac{100}{100} = 5
\]

The third step depends on whether the data set has top or bottom coding. If the data set has bottom and/or top coding those values are marked as swapped. Figure 4.1 shows how a data set with bottom/top coding is marked before the actual data swapping procedure.
The swapping procedure can be described with figure 4.2.

In the first step, the lowest un-swapped variable value is the first one. The range within which that first variable can be swapped is determined by \( P(a) \), since we chose \( P(a)=5 \) the range goes from \( a_2 \) to \( a_5 \). If the multivariate statistics deviate too much from the original the value/range of \( P(a) \) is too high and should be reduced.

Let’s assume that \( a_1 \), the value of variable \( a_1 \), is randomly swapped with \( a_3 \), the value of variable \( a_3 \). Then in the second step both \( a_1 \) and \( a_3 \) is marked as swapped, and the lowest un-swapped variable becomes \( a_2 \). Which means the value of variable \( a_2 \) can be swapped with any un-swapped variable value within the variable range \( a_3 \) to \( a_6 \). Let’s assume variable value \( a_2 \) is swapped with variable value \( a_5 \), and then both variables are marked as swapped in the next step.

In the third step the lowest un-swapped variable is \( a_4 \) and the range for swapping ends at \( a_8 \). It is evident by now that for each variable value that is swapped the range that the swapping can occur in is filled with swapped variables. Also, every new range, and with new we mean the iterative movement of the swapping range, can from start have some swapped variables. Since at least one new variable is included in the range for every swap step, every “lowest un-swapped” variable will have another variable to be swapped with. However, Moore (1996) does not mention how to deal with situations when the number of swap eligible variables is uneven, and one swap eligible variable is not swapped.

This procedure continues iteratively until all variables, or a pre-determined ratio, is swapped.
**Enhanced rank swap:** The enhanced rank is used on continuous variables, and it restricts the range for swaps, the result limits the distortion. The method is in many ways similar to rank swap as described above. However, this method focuses on finding a methodology that results in acceptable values for \( P(a), (Pb) \)...prior to the first swap. The aim is to keep the analytical value of the masked data but within an interval so that some summary statistics are preserved. That is, means of subsets that are swapped is preserved.

The first step is to determine which statistics to preserve, i.e. distributions. Moreover, both multivariate dependence/independence and means of subsets must be preserved. The objective of preserving the means of subsets originates from the aim to mitigate false conclusions about subsets in a perturbed data file. That is, the aim is to reduce the risk of drawing false conclusion about a subset. Suppose we have a data set where income is the swapping variable, if a researcher uses that data set to analyze women’s incomes the means and other univariate statistics can be misleading if the incomes of women are freely swapped with the incomes of men, and vice versa. The interval in which the swapping occurs in is assumed to be uniformly distributed.

To preserve the multivariate independence/independence one chooses an \( R_0 \) value so that \( 0.0 < R_0 < 1.0 \) and so that the expected correlation of two swapped fields \( (a \text{ and } b) \) equals \( R_0 \) times the original correlation of the fields \( a \) and \( b \). That is:

\[
E(R(a'b')) = R_0 \times E(R(ab))
\]
Where $R(a'b')$ is the correlation after the swap and $R(ab)$ is the correlation prior to the swap. $R_0$ is the expected value of the ratio of the largest covariance to the smallest. Using the equation above Moore (1996) concludes that one can find an acceptable estimate for $P(a)$ to preserve multivariate covariances with the following equation:

$$P(a) = 100 \times \frac{\sqrt{2 \times \text{var}(a) \times (1 - R_0)}}{(a_{\text{topcode}} - a_{\text{bottomcode}})}$$

When preserving the means of a subset, the aim is to swap two variable values so that $a_i$ (the randomly chosen variable to swap with) with $a_i$ (the lowest un-swapped variable that is going to be swapped), so that:

$$E\left[\frac{a_{i'} - a_i}{a_i}\right] = K_0$$

$$K_0 > 0$$

The value $K_0$ is the percentage alteration on each $a$ after the swap. If $K_0=0.2$, then the percentage alteration of $a$ is 20%.

To control the swap procedure so that each swapped value differs from its original value by $\pm K_0$. Moore argues that following equation gives a reasonable estimate for $P(a)$.

$$P(a) = 100 \times \frac{\sqrt{8 \times a}}{3 \left(a_{\text{topcode}} - a_{\text{bottomcode}}\right)}$$

Furthermore, Moore proves the relationship between $K_0$ and $R_0$ with following two equations:

1. $R_0 = 1 - \frac{4}{3} \times \frac{\bar{a}^2}{\text{var}(a)} \times K_0^2$
2. $K_0 = \sqrt{(1 - R_0) \times \frac{3}{4} \times \frac{\text{var}(a)}{\bar{a}}}$

So if $K_0$ is specified one can through equation 1 specify $R_0$, and vice versa. Assume $V(a)=1$ and $\bar{a}=1$, then it is evident that if $K_0$ is low $R_0$ will be closer to 1, and if $R_0$ is close to one $K_0$ will be
close to zero. That is, the distortion of the data set will be lower/higher based on the relationship between \( R_0 \) and \( K_0 \).

**A data-swapping technique using ranks:** Carlson and Salabasis (2002) introduced a new data swapping technique based on ranks and that can be applied on quantitative data. The basic idea behind this method is to exchange the values of a pre-specified subset of a large data set with the values of other subsets from the same larger data set. Assume that there is a dataset with two variables \( X \) and \( Y \). The first step is to randomly divide this data set into two equally large and disjoint subsets, \( M_1 \) and \( M_2 \), denote \( M_1=[X_1,Y_1] \) and \( M_2=[X_2,Y_2] \) and suppose \( M_1 \) and \( M_2 \) are ranked on \( X_1 \) (the values of \( Y \) will follow \( X \) after ranking). If the sample size is large enough, the values of \( X_1 \) should be approximately equal to the corresponding ordered values of \( X_2 \). The last step suggested by the authors is to swap the values of \( X_1 \) and \( X_2 \), the new data set can be denoted as \( M_1^*=[X_1,Y_2] \) and \( M_2^*=[X_2,Y_1] \) as the table shows below.

This technique can be also applied to multi-variable dataset. The authors illustrate a dataset with six variables as the figure 4.3. The figure is copied from Carlson and Salabasis (2002):

![Figure 4.3](image)

In this example, the whole dataset is divided into three subsets, \( M_1 \), \( M_2 \) and \( M_3 \), furthermore, three variables remain the same and variables \( Z \), \( U \) and \( V \) have been swapped. In these three data sets, \( M_1 \) works as the reference set, \( M_2 \) and \( M_3 \) work as the auxiliary sets, which means \( M_2 \) and \( M_3 \) will provide the values that replace the original values of \( M_1 \). The notations that appear in this figure describe the data swapping procedure for this method. For instance, the notation \( R_{Z1}^TR_{Z2} \) presents the process of how to exchange the values of \( Z_1 \) for those in \( Z_2 \).

The effect of this data swapping technique on the association between pairs of variables was
described later in this paper. The authors explore different cases with only two variables X and Y. They derive the cross product moments; however since the results are complicated they simulate studies under a bivariate normal distribution. The results of these simulations studies indicate that with the precondition that the added variables come from the same source, the more variables that are swapped, the higher the level of expected deterioration.

The authors’ research also shows how data swapping procedure affect the univariate statistics. They pointed out that the data swapping procedure can be considered as adding some kind of error or noise to the original source.

*Data swapping for protecting Census data:* To protect Census tables is more difficult than to protect sample-based data since census data contains whole population counts, while sample data presents only a part of the whole population.

The disclosure risks with census tables are small cells, i.e. small values, and cells with zeros that in itself does not reveal any information but can reveal together with other cells. The objective with statistical disclosure control on census data is to perturb small and zero cell values.

Shlomo et al (2010) introduced a data swapping technique to protect census data, called targeted rank swap. The proposed target variable for swapping is geography. The variable geography is divided into blocks, for instance a block can contain 4000 persons and 1900 households, or more or less.

In targeted data swap a sample of households is drawn and, by the statistical organization, a specified swapping rate (p) is chosen, and the households are swapped based on the variable geography. Put simply, households that are flagged as high risk i.e. found in small cells and have high disclosure risk, are swapped on the basis of the variable geography. E.g. municipality.

The determination of whether a household is high risk is based on whether an individual is high risk. An individual is high risk if she/he has a unique attribute. To determine if an individual is unique one calculates a risk score. Any variable that gets a risk score that is too high is flagged as risky; any variable that is unique is marked as unique. A variables that is unique for its geography block is swapped with a variable outside of its block. However, unique is not necessarily the same as high risk. Risk scores are calculated as follows:

- Calculate frequency counts for M selected key variables each having \( k_m \) \((m=1,\ldots,M)\) categories at the geographical level \( g \): \( N_{k_m}^g \) (not including individuals that have been imputed to account for the Census under coverage)
For every individual with values of categories \(k=(k_1, k_2, \ldots, k_m)\), calculate a score at each level of geography \(g\) by taking the average of the reciprocals of the counts:

\[
HR^g_k = \left( \frac{\sum_{m=1}^{M} 1/N^g_{k_m}}{M} \right)
\]

A threshold is set for each level of geography and those scores above the thresholds determine high-risk individuals.

High risk households are defined as any household having at least one high risk individual 

(shlomo et al, 2010)

Then one can either swap in geographical areas that are small or swap a percentage of the high-risk households in a geographical area. In other words, a sample of households to swap is selected. These households are then paired with other households based on a set of control variables. Then the geographical variable of those households are swapped. Control variables can be age, sex, and ethnicity. Moreover, households are swapped within a certain geographical distance, so that someone who lives in “Piteå” does not appear to work in “Malmö”. This also ensures less bias in the data set. Using geography as the basis for swapping rather than age provides a data set with less bias. That is, if the swapping variable is age the records can show a five year old as married with children, which in turn make the record illogical and inconsistent. This method preserves marginal distributions and introduces ambiguity and uncertainty to a potential data intruder/snooper as to whether identification is correct. A higher \(p\) protects more cells than a lower \(p\), however, as in all SDC methods one has to choose between utility versus information loss. The advantages with this method are that the marginal distributions are preserved and that it is easy to use.

5. How do statistical organizations work?

There is no standard method that all statistical organizations use. The usage depends mainly on legislation, and since no country has the same laws no country uses the same methods. Often organizations tailor their own methods based on the theoretical (and practical) methods mentioned above. One of the main implications of legislation is what an organization can publish and how this inflicts on the publication form, aggregated or micro data.

The data swapping method used in the United Kingdom, the UK office for national census of England and Wales, Scotland and Northern Ireland, is a modeled variant of targeted data swap...
for census data. Their aim is to maximize output utility and minimize information loss while at the same time protect against disclosure. The method can be used to protect individuals living in a communal establishment, that is prisons, hospital etc. from disclosure. The choice of method is based on a need of perturbed data that both keeps respondents confidentiality and yet maintains as much data utility as possible. The chosen method, unlike suppression, did not miss any values and had been used before. When a similar method was used there were some concerns that the public would not perceive that confidentiality was protected.

The UK census, unlike the Swedish census, is not based on registers but depends on responses from the public. This in turn means that the census data might have non-response rates that in themselves are a protection of the data set.

In the case of communal establishments individuals swap residency as an option to the geographical variable of the communal establishment. That is, since there might only be one prison in one community, a prisons geographical attribute cannot be swapped without distorting the data too much; hence individuals residing in that prison are swapped instead.

The Australian Bureau of Statistics (ABS) has the right to collect statistical information and this authority is given by national legislation. The basic requirement of this legislation states that the data which seems to enable identification of a particular person or organization is not allowed to be released. Given the legislation Australia adapts its SDC-methods to make sure that the released data do not have any disclosure risk.

The Census of Population and Housing is conducted by the ABS every five years. With the aim of protecting Census data, a new method was developed by Fraser and Wooton (2005). This new method is intended to be used within a web-based product that allows users to define their own tables, and has the ability to protect the Census data automatically. Shortly the method they practice can be described as: they use a permanent numeric key to each unit in a record. Consistent values are generated with the keys. The generated values are applied in the cells of a table. Furthermore, the ABS does not intend to change their method for protect Census data which means this method will be used in the following years.

Statistics Sweden does not use methods to mask micro data. However, methods to mask tabular data are used. In case of micro data they use remote access, MONA, as a way to technically safeguard micro data. MONA (micro data online access) stores micro data files on a protected data bank. Researches with permit can access the files with an Internet connection. That is, the database contains micro data files and a researcher can, with permission, use the database to do research. However, the researchers do not get direct access to the micro data, only the output. To
protect data, they sometimes use controls on the output, and demands on output data is examined so that sensitive information is not released.

Furthermore, researchers can access micro data; however, legislation on secrecy follows the micro data, hence the responsibility of protection as well. So any researcher that gets access to micro data from a statistical organization must follow the same laws of secrecy as the statistical organizations.

However, Eurostat has asked for census from all European countries and since Eurostat does not mask data themselves, each country must mask data according to their own legislation. In this case, statistics Sweden has appointed a project group to accessed register data from 2011 and formulate a method of masking micro data, so that the aggregated data is masked, and consistent.

The data that statistics Sweden is going to provide what they call hyper cubes, tables with millions of cells. These tables consist of many variables, and given Sweden’s small population, many cells are either blank or have small values. This in turn means that statistics Sweden, according to Swedish legislation, must consider ways to protect data. Their main goal is to publish protected data that is consistent over all tables. Given that tabular protection would be tedious and take tremendous time, they try a method of first masking the micro data, and then publishing aggregated tabular data that is masked.

We understand that the project group will propose data swap. Based on their need of consistent masked data, cost effectiveness and a method that works on census data, researching how other organizations have done, data swapping seemed to not only meet their demands but was also a method that have been used and approved by others. The data swapping method they use is tailored for their needs. They use variant of the targeted record swap where the swapping variable is geography. They use a specially coded program to conduct the swaps.

6. Discussion and summary

Today it is easier to access computers, Internet and data than before. Also the amount of information and registers that can be found and that is published is larger than before. Hence, using information that is available makes disclosure easier. Furthermore, consumers and private users’ demand more detailed information from statistical agencies. The accessibility of technology and information is a problem for the publications of statistical information, and it limits the publications when having confidentiality in mind.

Although there are many statistical disclosure methods each statistical agencies must deal with
statistical data depending on the data, the form of the data and the variables. That is, there is no clear-cut answer to which method to use. Decisions are based on type of data, how it is going to be published and an evaluation of what variables are at risk of disclosure which means that statistical organizations must be flexible in their practice of SDC-methods.

One of the advantages with micro data is that it has analytical advantages to aggregated data; however there are risks associated with micro data as well. Since micro data usually contains more variables, both direct and indirect identifiers, than aggregated data, the disclosure risk poses a higher chance of identification. Comparing the available micro data SDC-methods we see no clear answer to which method should be used. However, some data swapping techniques seem to have advantage over other SDC-techniques. One advantage is that no information is removed; hence the information loss is relatively smaller. Another advantage is that no new elements are introduced, so the analytical value is relatively higher. Given the use of internet and other techniques, data swapping seems to be the method that best can meet the demands of today.

Some of the outstanding issues are the growing use of social medias, such as Facebook, and the information there that can lead to disclosure of individuals or businesses. On other hand, how much should factors like that affect statistical organizations risk assessment? That is, how much time should be consumed on analyzing possible attributes that individuals might or might not post on the Internet that can lead to disclosure?
7. References


